NCBI HMM Collection (NCBIfam) Release Notes (NCBIfam)

Version 1.1, November 27, 2017

This document serves as the set of release notes for release 1.1 of all components of NCBIfam. In the future, new releases of the various components may not be synchronized, and each component may have its own release notes.

The introduction of NCBIfam is discussed as a new resource for rule-based annotation in the article by Haft, et al., *RefSeq: an update on prokaryotic genome annotation and curation*, for the January 2018 database issue of *Nucleic Acids Res.* Please cite this article, PMID: 29112715, as necessary, if no more recent article on the NCBIfam collection has appeared.

A protein profile hidden Markov model (HMM) is a statistical model that generates scores of various types for any protein sequence it encounters. A protein that meets all required score thresholds is said to be "hit" by the HMM, and is classified as a member of the family of proteins that the HMM helps to define. In this collection, all HMMs are built from multiple sequence alignments, using the HMMER3 package. When an HMM hits a protein, that hit may be sufficient information for deciding that the protein should be annotated in various ways: protein name, gene symbol, enzyme commission (EC) number, etc. Most models in this release of NCBIfam include a name that can be applied during automated annotation to any matching protein, as long as no other evidence with higher precedence exists to overrule it.

Divisions

- **AMR HMMs**
- **PRK HMMs**
- **other HMMs**
- **NCBI name revisions for TIGRFAMs**

General Notes on NCBIfams.

This HMM library does not distinguish gathering thresholds from trusted cutoffs. Identical values are used for both, so the command line arguments --cut_tc and --cut_ga will behave the same. For most models, noise cutoffs are not set, so the argument --cut_nc should not be used.

In this incremental release of NCBIfams, release 1.1, several metadata fields we are collecting, including Enzyme Commision (EC) number, gene symbol, and literature citations, are not yet provided. These data will be included in future releases.

HMMs are distributed in two forms. In one form, the HMMs from a set are concatenated into an HMM library, with the filename suffix .LIB. Users will need to run the utility hmmpress from the HMMER3 package to create the indexes needed to scan a sequence

against the library, using hmmscan (search one sequence at a time against a library of HMMs). The HMMs are also distributed as a compressed archive.

NCBIfams are used in NCBI's Prokaryotic Genome Annotation Pipeline (PGAP), and are used in conjunction with HMMs imported from the Pfam [PMID:26673716] and TIGRFAMs [PMID:23197656] collections. Starting with PGAP version 4.1, the genome annotation pipeline tests all possible ORFs for hits from HMMs during structural annotation, and forces creation of a gene feature if an HMM hit indicates the predicted protein is a member of defined protein family. During preparation to use HMMs to support structural annotation, we flagged a number HMMs as inappropriate to trust for structural annotation and send feedback to the source database.

All names we provide as curated protein product names associated with HMMs have been tested for acceptability in the submission of sequence data to GenBank. Some names may continue to generate warnings as both naming standards and name validation software are updated, but every curated name provided should be acceptable.

We have assigned many HMMs to a specific *family type*, a field that controls the relative precedence of an HMM in automated annotation pipelines. The meanings are as follows:

exception – the most specific type of HMM, able to overrule a name assignment from an equivalog-level (or lower) HMM. Models of this type may be used to contradict the annotations an HMM of lower precedence might make, or simply to provide additional information about a subgroup with a notable extra characteristic.

equivalog – the relationship of full-length protein homologs that share a specific function, by virtue of conservation of function from a shared common ancestral protein. **equivalog_domain** is an independently functioning region, with specific function, that often shows up in longer, multifunctional proteins.

subfamily – the homology relationship shared by groups of proteins with full-length homology but with variable functions. The name given to a subfamily HMM for use in annotation should apply to all members of the family, not to one specific member. Subfamily models usually describe and separate a notable subgroup from a larger domain or superfamily collection.They do not include all homologs out to the limits of detection. A **subfamily_domain** is such a region that regularly occurs in the context of variable domain architecture. A **paralog** family is a type of subfamily that was observed to have a large number of members in some model genome. In general, such a model is expected to hit additional proteins in other genomes, although without any guarantee to find the full extent of the paralogous family in those other genomes.

superfamily - the relationship among a large set of proteins, stretching to the limits of homology detection, but finding primarily proteins that are consistent in their domain architectures. This family type is now rarely assigned, since another type usually is more apt.

domain – a region of sequence homology that can be shared by proteins that lack homology elsewhere, and that is not responsible for a specific autonomous function.

repeat – similar to a domain, but usually shorter, and usually found in two or more tandem copies when they occur in a protein at all.

AMRfam collection.

The collection of HMMs in NCBIfam that was built for analysis of AntiMicrobial Resistance (AMR) genes (and the proteins they encode) is called **AMRfam**. This set is designed to identify acquired AMR genes such as those borne on plasmids or in transposons, and also intrinsic genes such as chromosomally encoded beta-lactamases. AMRfam does not currently support of analysis of adaptive resistance mechanisms such as point mutations to drug targets or loss-of-function mutations in transporters. In the current release, 544 models are used to analyze protein families considered reportable in the Pathogen Detection system Isolates Browser, https://www.ncbi.nlm.nih.gov/pathogens/isolates. Thirty additional HMMs in AMRfam describe intrinsic proteins that contribute to resistance, typically as multidrug export proteins, but that tend to be insufficient by themselves to actually confer resistance. Consequently, we provide two versions of the library of AMR HMMs. The smaller library consists exclusively of those HMMs required for our AMRFinder utility to identify AMR genes reported in the isolates browser (a distributable version of the utility is planned, but not yet available). The larger library contains additional HMMs for the "non-reportable" protein families that we detect, but do not report in the isolates browser.

Because the collection of AMR HMMs is hierarchical, a single protein can have several HMMs that score it above their respective scoring thresholds. In nearly every case, the HMM that is the most specific for a protein is the one that gives the highest score. Note, however, that AMRFinder itself uses information from an explicit tree of AMR protein family relationships rather than this simple heuristic. A family type is provided explicitly for many models, but the model's position according to the tree supersedes the precedence information that the family type can provide.

Sources used to develop the sets of AMR protein families for which models were built were numerous. They include Lahey Clinic and Pasteur Institute compilations of beta-lactamases and Qnr proteins (http://www.lahey.org/studies/ and http://bigsdb.pasteur.fr/klebsiella/klebsiella.html), the Comprehensive Antimicrobial Resistance Database (CARD) [PMID:23650175], ResFams [PMID:25003965], ResFinder [PMID:22782487], personal communications from Drs. Marilyn Roberts, Derrick Crook, Shaohua Zhao, and Sally Partridge, and submitters using NCBI's beta-lactamase allele submission portal, along with extensive review of the primary and review literature.

PRKfams

We began with 7471 clusters of prokaryotic proteins ("PRK clusters"), of "reviewed", "curated", or "provisional" status, from NCBI's ProtClustDB database, and revisited annotation of the protein name field. The level of sequence diversity within each cluster differed greatly, which complicated efforts to assign appropriate cutoff scores. Therefore, clusters exhibiting high sequence diversity were broken up into two or more smaller clusters with similar levels of conservation in their multiple sequence alignments. HMMER3 (release 3.1b2) hidden Markov models (HMMs) were constructed from the multiple alignments, using hmmbuild. Heuristics used to estimate HMM cutoff scores were based on seed alignment length and diversity, scores obtained for members of the PRK cluster itself, scores from other PRK clusters with different names. When a single PRK cluster becomes the source of multiple HMMs, a single protein may be hit by several of them, but overlapping hits from HMMs from different PRK clusters are rare.

PRK HMM curation was reviewed by comparison to data from other sources, including TIGRFAMs, Pfam, SwissProt, GeneRIF, and PDB.

PRK HMMs have not yet been grouped by the biosystems to which their member proteins belong. Once biosystems are defined for a subset of PRK clusters, comparative genomics and biosystem reconstructions will be used to further refine cutoff scores of the derived HMMs.

In this release, all PRK models are treated as equivalog, but that data field typically is not populated explicitly in the PRK HMM library metadata file.

General (non-AMR, non-PRK) HMMs

This portion of NCBIfams consists of models built for a wide variety of purposes, but most to provide new coverage of "dark matter", the part of the world of protein sequences that no prior set of tools was describing effectively. These HMMs may solve problems noted in structural annotation (gene-finding is especially hard for protein shorter than 60 amino acids), functional annotation, or both. Several HMMs were built to describe signature regions (usually repeat regions) in sets of proteins that exceeded 3000 amino acids in length yet had no HMM hits anywhere along their length to confirm. Note that the longest few proteins in a bacterial pathogen often play a role in virulence, and the lack of tools to find, classify and describe them can be a barrier to their characterization and understanding.

NCBI name field curation for TIGRFAMs models

NCBI will begin to host TIGRFAMs within the next year. In the meantime, we have been using TIGRFAMs HMMs in PGAP annotation. We have reviewed a substantial fraction of TIGRFAMs models, both for name format and for updates to functional names based on the latest literature. For 3433 HMMs from TIGRFAMs, we have either ratified the name from TIGRFAMs release 15.0 for use in automatic annotation pipelines, or supplied an updated

name. These approved names appear in the column labeled “product_name” in the file TIGRFAMs.tsv.

All names appearing in the “product_name” field in TIGRFAMs.tsv have been deemed suitable for use in submission of data to GenBank. Note that some names from the original source database (TIGRFAMs release 15.0 from the J. Craig Venter Institute) have features with their names, such as inclusion of genus and species identifiers, that complicate submission to GenBank. Those who already use TIGRFAMs to help prepare data for submission to sequence databases should use the updated “product_name” field supplied in TIGRFAMs.tsv rather than older names associated with those same models.

NCBIfam release dates and statistics

2017-Jun-01 Release 1.0
2017-Nov-27 Release 1.1

```
NCBIfam-PRK HMMs
Release 1.0     11498 models (6331 re-curated) from 7472 clusters
Release 1.1     11497 models (7355 re-curated) from 7472 clusters


NCBIfam-AMR HMMs
Release 1.0     558 models (528 used by AMRFinder)
Release 1.1     574 models (544 used by AMRFinder)


NCBIfam-gen HMMs
Release 1.0     110 models
Release 1.1     184 models
```

For questions or comments about NCBIfam, please contact pd-help@ncbi.nlm.nih.gov.